Department of
AUTOMATIC
CONTROL
and SYSTEMS
ENGINEERING
UNIVERSITY of SHEFFIELD

MODELLING THE FLOW INHERENT IN SPEECH REPRESENTATIONS

L Baghai-Ravary*, M O Tokhi* and S W Beet**

* Department of Automatic Control and Systems Engineering,
** Department of Electronic and Electrical Engineering,
The University of Sheffield, The University of Sheffield,
P O Box 600, Mappin Street, Sheffield, S1 4DU, UK.

Tel: (0742) 825136.
Fax: (0742) 731 729.
E-mail: O.Tokhi@sheffield.ac.uk.

Research Report No. 551

November 1994

Abstract

This paper presents two new methods for modelling the flow inherent in speech: flow-based prediction (FBP) and acoustic flow interpolation (AFI). These are presented as extensions of the form of prediction implied in calculating the delta and delta-delta coefficients often used in automatic speech recognition. All these methods are presented as special cases of a general vector linear prediction model, but it is shown that the new techniques, which make the flow of features within the data explicit, are significantly better at modelling spectrogram-like data.

Several speech representations, using both parameteric and non-parametric analyses, are discussed both in terms of their ability to represent speech accurately, and of their appropriateness to these flow-based models. AFI and FBP error coefficients, for both male and female speakers, are measured and compared with the delta and delta-delta coefficients. Wherever possible, the parameters and methods used to produce the representations have been chosen to be directly comparable with one another.

Key words: Acoustic flow interpolation, flow-based prediction, acoustic flow modelling, delta coefficients, delta-delta coefficients.

CONTENTS

1 Introduction

Most of the articulators involved in speech production are not able to move abruptly. Speech signals can, therefore, be considered piecewise-continuous, except, for example, during plosives (where the signal statistics change rapidly). Plosive sounds have a very short duration, thus, those segments of speech which do not fit the piecewise-continuous model are only a tiny proportion of the overall speech duration, and most speech evolves smoothly with respect to time. Such forms of nonstationarity are more explicit in some speech representations than others, due to the different structures which they reveal.

One common feature of spectrogram-like representations of speech is that changes in formant frequencies or pitch will result in features in the observation vectors "migrating" from one element to another. This is illustrated in Figures 1 and 2, which show the spectrogram of a short phrase and the flow of the features within it (the "spectrographic acoustic flow") respectively. Current automatic speech recognition (ASR) systems are unable to deal with this form of evolution, and can only treat such gradual changes as a sequence of unrelated steps. This requires significant temporal over-sampling to ensure that consecutive frames with smoothly-evolving characteristics can be identified as such by the small Euclidean distances between them.

Some ASR systems can use such identification of smoothly-evolving segments to improve recognition performance. This is most often done via delta coefficients (Shirai and Mano, 1986). Alternatively, delta-delta coefficients can be used, or a full state-dependent first order linear predictor can be introduced into the recognition process (Kenny et-al, 1990). Delta and delta-delta coefficients can also be thought of as prediction errors, but in all these cases, the implied predictor is still stationary over a time-scale which often includes significant changes in formant frequency and even pitch.

Recognisers based on hidden Markov models (HMMs) can use triphone and diphone models to cope with the dynamics of the signal, but still, large quantities of training data, HMM states and computational power are required. The worst of these problems is the need for instantiations of every possible triphone or diphone during training. Thus, at present, speech recognition (and also coding) systems do not fully account for speech dynamics, requiring significant temporal over-sampling and even then, attributing undue importance to many highly redundant parts of the signal. This paper presents an investigation into the development of two new methods of modelling speech dynamics, namely, flow-based prediction (FBP) and acoustic flow interpolation (AFI). These will form the basis of subsequent investigations at developing robust automatic speech recognition systems.

2 Flow models

Flow-based prediction is a new method to deal with the dynamics of speech. This method characterises the level of redundancy in the data by tracking the features within the speech signal, and predicting both their flow and their values. Flow-based prediction requires no training data, no increase in the number of HMM states and not much computation. Acoustic flow interpolation is an alternative formulation of the same model, but, which performs interpolation rather than prediction. This makes it more robust, especially during any abrupt transitions from one continuous segment of speech to the next.

This section describes the new methods and explains how delta coefficients and delta-delta coefficients can be viewed as errors of zeroth and first order linear predictors, respectively.

2.1 Vector linear prediction

Figure 3 shows a zero-order linear prediction process where, $M_n(i)$ and $M_{n-1}(i)$ represent the values of element i of frames n and $n-1$, respectively. The delta coefficients are produced when the n^{th} frame is predicted by the preceding, $(n-1)^{th}$ frame. This predictor only allows for horizontal transitions and assumes that the signal is stationary so that each frame is equal to its predecessor;

$$M_n(i) = M_{n-1}(i) \tag{1}$$

To produce the delta-delta coefficients, a first order linear predictor with a coefficient matrix equal to the identity matrix is used. This process is shown in Figure 4, where x represents the difference in value between element i of frame M_n and that of its predecessor. This predictor uses the $(n-2)^{th}$ and $(n-1)^{th}$ frames to predict the n^{th} frame. It assumes that the difference between the previous two frames remains constant;

$$x = M_{n-2}(i) - M_{n-1}(i) = M_{n-1}(i) - M_n(i) \tag{2}$$

Thus, simplifying equation (2) yields the n^{th} frame as

$$M_n(i) = 2M_{n-1}(i) - M_{n-2}(i) \tag{3}$$

2.2 *Flow prediction*

The fundamental assumption in both the AFI and the FBP models is that the successive speech segments are correlated and the flow inherent in speech is constant. This means that the features flowing in a particular direction would continue in that direction without exceeding the frame boundaries. In this manner, the n^{th} frame is predicted using two frames, n_1 and n_2, from a sequence. The relative positions of n_1 and n_2 in the sequence differ between the AFI and the FBP methods, but both operate by calculating the most likely flow between the n^{th} frame and its immediate predecessor.

FBP is an extended form of linear prediction. The prediction process is described in Figure 5, where x and y represent the difference between the values and positions, respectively, of the i^{th} element of frame M_{n-2} and the j^{th} element of frame M_{n-1};

$$\begin{aligned} x &= M_{n-1}(i) - M_{n-2}(i) \\ y &= j - 1 \end{aligned} \tag{4}$$

The n^{th} frame is predicted from the previous two frames, and the direction in which the features flow, is also tracked. Dynamic programming can be used to match the elements of M_{n-1} optimally with the elements of M_n so as to minimise the mean square

This is like replacing the first order predictor by a zero order predictor, but retaining the flow model, as the j^{th} element is still transferred to the k^{th} element.

In AFI, the n^{th} frame is estimated using the preceding, $(n-1)^{th}$, and the following, $(n+1)^{th}$, frames. This is described in Figure 6. Using dynamic programming, links between M_{n-1} and M_n are made, and it is assumed that

$$M_n(j) = \frac{M_{n+1}(k) + M_{n-1}(i)}{2} \tag{11}$$

where, $j = \dfrac{k+i}{2}$.

The local distance matrix is defined in terms of all the possible links between frames M_{n-1} and M_n which can be extrapolated to end at elements of M_{n+1}. Any links which do not satisfy that condition are disallowed by setting the corresponding local distances to ∞. Elsewhere, the values are assumed to evolve linearly along the selected links and, thus, the squared differences between frames M_{n-1} and M_n are used;

$$d_{AFI}(i,j) = \begin{cases} \left[M_n(j) - M_{n-1}(i)\right]^2 = \left[\dfrac{M_{n+1}(k) - M_{n-1}(i)}{2}\right]^2 = \dfrac{1}{4}\left[M_{n+1}(2j-i) - M_{n-1}(i)\right]^2 & \text{for } 1 \le k \le P \\ \infty & \text{otherwise} \end{cases} \tag{12}$$

3 Difference coefficients

The difference coefficients, which are the delta, delta-delta, FBP error and AFI error coefficients, are the differences between the actual signal and that predicted by the respective prediction algorithms. The mean square error (MSE) over each frame is calculated, and this is normalised to give an error score;

$$\text{Error Score} = \sqrt{\frac{MSE}{MSE + (\text{within frame variance})}} \tag{13}$$

This formulation is used so that the different speech representations, which all have different means and variances, can be compared fairly. The average error score for a whole sentence is then

$$\text{Average Score} = \frac{1}{N}\sum_{1}^{N} \text{Error Score} \times 100\% \tag{14}$$

where N is the number of frames in the sentence. This score is calculated, later for each of the pre-processors. In this manner, a value of 71% implies that the MSE is as "big" as the signal, while 50% indicates the representation's variance is three times bigger than the MSE.

4 Processing considerations

An aim of this paper is to allow valid comparisons to be made between the different representations of speech. Thus, wherever possible, the same parameters and methods are used. This section describes the common parts of the pre-processors.

4.1 Pre-emphasis

The data used here was taken from the TIMIT database, sampled at 16,000 samples per second. The speech was pre-emphasised, giving roughly 6dB per octave gain above 500Hz, prior to each analysis. This is a generally accepted method for improving speech representations, although there is some doubt about its appropriateness for recognition of vowels. Most of the acoustic cues for vowel sounds lie below about 3kHz, and Paliwal (1984) has found that pre-emphasis puts undue weight on higher frequency components resulting in deterioration of vowel recognition by approximately 1%. Speech is constructed of both vowels and consonants, and in the case of consonants the acoustic cues are normally found in the higher frequency regions where pre-emphasis is beneficial. For this reason, the pre-emphasis applied here was slightly less extreme than most. It was based on

The solution to the relations in equation (23) can be found efficiently via the "reflection coefficients", k_i, of an equivalent lattice filter as shown in Figure 7. In this manner, successive values of k_i can be found by setting

$$\begin{aligned} f_0(n) &= g_0(n) = x(n) \\ f_i(n) &= f_{i-1}(n) - k_i g_{i-1}(n-1) \\ g_i(n) &= g_{i-1}(n-1) - k_i f_{i-1}(n) \\ E_i &= \sum_{n=p}^{N-1}\left[\left(f_i(n)\right)^2 + \left(g_i(n)\right)^2\right] \end{aligned} \tag{24}$$

which yields

$$f(n) = f_p(n), \quad g(n-p) = g_p(n) \quad \text{and} \quad E = E_p \tag{25}$$

and minimising E_i with respect to k_i for each stage of the filter, yielding k_i as

$$k_i = \frac{\sum\limits_{n=i}^{N-1} f_{i-1}(n) g_{i-1}(n-1)}{\frac{1}{2}\sum\limits_{n=i}^{N-1}\left[\left(f_{i-1}(n)\right)^2 + \left(g_{i-1}(n)\right)^2\right]} \tag{26}$$

Finally, if required, the reflection coefficients can be converted to the AR model coefficients (the "a" coefficients mentioned in equation (22) earlier). The discrete-time transfer function, $H(z)$, for this lattice filter is, thus,

$$H(z) = \frac{F(z)}{X(z)} = 1 - \sum_{m=1}^{P} a(m) z^{-m} \tag{27}$$

Thus, if the k_i values are reinserted into equation (24), the transfer function for each stage of the lattice filter can be calculated recursively and equation (23) solved, yielding the values of the coefficients a.

4.4 Autocorrelation functions

The autocorrelation function for the Blackman-Tukey power spectral density (PSD) estimate was estimated from the inverse discrete Fourier transform of a periodogram (described later in section 6.1), and windowed with the autocorrelation function of a minimum 4-sample Blackman-Harris window (Elliott, 1987). This in itself is a finite-duration function, non-negative for all time and at all frequencies. It is, therefore, a valid window and also yields a valid (non-negative) PSD estimate. It has similar advantages over other window functions to those mentioned later in section 5.2.

4.5 Logarithms

Power spectrum estimates are normally encoded on a log scale. In this paper, this scale is approximated by a function with similar, but more well-behaved, numerical properties. The same function is used to encode the vocal tract area functions, and in the intermediate calculations for the cepstrum.

In practice, log scales can cause problems when numbers become very small, and are totally impractical if numbers can become negative (due to rounding errors, etc.). To avoid this, it is usual to set a lower threshold on the data values, before the log is taken. This, however, assumes that the range of values is known a priori. To avoid having to estimate the respective ranges, a log scale can be approximated by taking the N^{th} root of the data value;

$$\ln(x) = N\left(\sqrt[N]{x} - 1\right) \tag{28}$$

Here, N is a constant defining the range over which equation (28) is valid. The larger N is, the wider the range on either side of $x = 1$, for which the approximation holds. Furthermore, if N is chosen to be a positive, odd integer, equation (28) will be monotonic and calculable for any real value of x.

In the data presented here, a value of $N = 5$ was used, giving an effective dynamic range of 200:1 regardless of the mean level of the data. Interestingly, this value is similar to that used in many auditory models to simulate the dynamic range compression of the human ear.

4.6 Auditory frequency scales

The pre-processors that produce a PSD estimate, or use one at an intermediate stage of their computation (i.e. the cepstrum, described later in section 6.5), can be subjected to frequency-warping, to match the frequency resolution and scaling of the human auditory system. This has been found to improve speech recognition and coding performance (Makhoul and Cosell, 1976).

The most common form of frequency warping is according to the "mel scale", an early approximation to the "natural" frequency scale of human perception (Deller et. al, 1993). However, the experiments used to measure this scale give no direct indication of frequency resolution; they only indicate the listeners' opinions as to the ratio between the frequencies of two tones. Because of the experimental procedures used, and the difficulty of accurately perceiving pitch (especially at very low or very high frequencies), the mel scale can only be treated as an approximation to a true "perceptual frequency scale". In reality, such a true scale is improbable since perception is a very complex and non-linear process, and, thus, one would expect such a frequency scale to vary, depending on the level and spectral composition of the signal in question.

The warping function used here was, therefore, chosen to be the equivalent rectangular bandwidth (ERB) scale of Moore and Glasberg (1987). This was developed specifically to characterise frequency resolution in a power spectrum model of human perception (Fletcher, 1940). Moore and Glasberg measured the frequency resolution of the ear using a test signal with a broad-band noise masker. In addition to the effective filter bandwidth of human perception at different frequencies, this procedure yields a modified frequency scale which maintains a constant perceptual ERB at all frequencies;

$$\begin{gathered}\text{ERB (Hz)} = 6.23F^2 + 93.4F + 28.5 \\ \text{Constant ERB} - \text{rate frequency warping: } F \rightarrow 11.17\ln\left(\frac{F+0.312}{F+14.765}\right) + 43.0\end{gathered} \tag{29}$$

The auditory representations considered here have, therefore, been processed to exhibit the ERB frequency resolution by forming a weighted sum of the periodogram PSD estimate for each of a set of equally spaced points on the constant ERB-rate scale. The weighting is triangular, with unit height and width of one ERB (see Figure 8). This form of weighting is like that traditionally used for mel-scale processing, and was used in preference to the one suggested by Moore and Glasberg (1987) purely for the sake of simplicity; the exact form of the weighting suggested by Moore and Glasberg (1987) changes depending on the signal level and involves more complicated (and thus, time-consuming) calculations.

5 Representations of speech

The pre-processors used in this investigation are discussed in this section. Both the male and the female sentences, "She had your dark suit in greasy wash water all year", which form the basis of this study, were taken from the TIMIT database. Diagrammatic representations of all the male processed speech are also included and discussed.

Auditory modelling is defined in the frequency domain. Thus, the parametric methods which do not involve calculations in the frequency domain have not been subjected to the auditory transformation.

5.1 Periodogram

This is the most common method for visualising speech signals. It is formed by taking the DFT of a windowed segment of speech, and finding the modulus squared of each complex output value;

$$\begin{gathered}\text{Discrete Fourier transform} = X(m) = \sum_{n=0}^{N-1} x(n)h(n)\exp\left(-j2\pi\frac{nm}{N}\right) \\ \text{Periodogram} = |X(m)|^2\end{gathered} \tag{30}$$

where, $x(n)$ and $h(n)$ are the speech and window sequences, each of length N, respectively. In this manner, it provides an estimate of the PSD which is only degraded by the spectral effects of temporal windowing. The frequency resolution of the periodogram is inversely proportional to the length of the input frame (for a given window shape), and cannot be controlled independently, except by changing the window.

The choice of window is restricted by the expected dynamic range of the elements in each PSD estimate, and the required degree of temporal continuity. To give temporal continuity across pitch-pulses, with adult male speech, this method can only give a narrow-band spectrogram, clearly resolving individual pitch harmonics. Figure 9 show such periodograms of the male sentence, but the visual resolution of these diagrams is not sufficient to show the true frequency resolution of the data. A typical periodogram and auditory periodogram frame (during voiced speech) are, therefore, shown in Figure 10. This reveals far more detail than could be seen in Figure 9. Such data has a complicated flow structure, since the frequency of the pitch harmonics may, for example, rise, while the formants are falling.

As with most of the representations presented here, the most obvious difference between the periodogram and its auditory version, is the frequency warping; at low frequencies, the auditory representation is stretched, while at high frequencies, it is severely compressed. In this case, however, the changing bandwidth of the auditory representation has a further effect on the images (and the underlying data). At low frequencies, where the auditory bandwidths are small, individual pitch harmonics are clearly visible. However, even when examining the single-frame plot in Figure 10, the pitch harmonics have been suppressed at high frequencies. It should be noted that, despite the overlapping of the data windows, there is still very slight evidence of temporal modulation when the pitch period is low.

5.2 Blackman-Tukey power spectrum

One method for controlling the resolution of a periodogram is to window an estimate of the autocorrelation function, rather than the data itself. This allows the frequency resolution to be reduced without losing temporal continuity. However, for negative PSD estimates to be avoided, the window must have a non-negative Fourier transform. In the work reported here, the window was made equal to the autocorrelation function of a suitable prototype window, simultaneously ensuring that the created window has finite duration and non-negative Fourier transform;

$$\begin{gathered}\text{Autocorrelation function} = A_x(m) = \sum_{n=0}^{N-1}\left[x(n)h(n)x(n+m)h(n+m)\right] \\ \text{Window} = A_h(m) = \sum_{n=0}^{N-1} h(n)h(n+m) \\ \text{Blackman - Tukey PSD estimate} = \text{DFT}\left(A_x(m)A_h(m)\right)\end{gathered} \tag{31}$$

This method can, thus, give a broad-band spectrogram, characterising formant structure rather than pitch. The particular window used here gave a spectral resolution of 400Hz, to ensure that even the highest female pitch was suppressed. However, since the resonances of the vocal tract are generally narrower than this (Deller et. al, 1993), the resulting spectrum is somewhat blurred. Furthermore, it can be seen in Figure 11 that the auditory transformation has not had much effect, since most auditory filters have bandwidths less than the 400Hz "blurring" introduced by the autocorrelation window. It should be noted that, again, there is slight evidence of temporal modulation when the pitch period is low.

5.3 Maximum entropy power spectrum

The power spectrum of an AR process can be obtained by calculating the parameters of the AR model from the autocorrelation function of the signal. This is done using the Burg's method. The maximum entropy method (MEM) PSD estimate is then obtained by multiplying the innovation power by the transfer function of the implied recursive filter;

$$\begin{aligned}\text{Innovation power} &= \sigma^2 \\ \text{AR transfer function} = H_{AR}(f) &= \frac{1}{1+\sum_{m=1}^{p} a_m \exp\left(-j2\pi\frac{f}{f_s}m\right)} \\ \text{Maximum entropy PSD} &= \sigma^2\left|H_{AR}(f)\right|^2\end{aligned} \qquad (32)$$

where, f_s is the sampling frequency.

Since speech can only be loosely approximated as an AR process, the resulting PSD estimate tends to be a bit too "peaky", with occasional false peaks ("peak splitting"). Nonetheless, as seen in Figure 12, it can make formant tracks very clear, both in the conventional and the auditory versions.

5.4 Maximum likelihood power spectrum

This is variously referred to as the minimum variance PSD estimate, the maximum likelihood method (MLM) or Capon's method. It involves the design of an FIR filter for each frequency where an estimate of the PSD is required. These filters have unity gain at the design frequency, but with minimal output power. The technique attempts to attenuate all but the frequency component of interest, and can be considered as a data-adaptive DFT.

The order of the filters determines the maximum number of frequency regions which can be attenuated, and is chosen according to the application. To resolve formant structure while suppressing pitch information, the filter order should be chosen to be slightly more than twice the maximum number of formants, as in linear prediction analysis.

The power from each filter is calculated from the AR model for the signal, without explicitly implementing the filters (Musicus, 1985);

$$\text{Maximum likelihood PSD } = \frac{\sigma^2}{\sum_{k=-p}^{p} \mu(k)\exp\left(j2\pi\frac{f}{f_s}k\right)} \qquad (33)$$

where, $\mu(k) = \sum_{m=0}^{p-k} (N+1-|k|-2i)a_m a_{m+|k|}$ and $a_0 = 1$.

Figure 13 shows that the MLM method generally has frequency resolution comparable with, or slightly better than, that of the Blackman-Tukey method, and with superior temporal continuity. The latter effect is due to the Blackman-Tukey method's need for data windowing, which is avoided in the maximum likelihood approach, being based (in this example) on Burg's method of AR parameter estimation.

5.5 Cepstrum

Since speech can be considered as the product of a source spectrum and a vocal tract transfer function, pitch information can be separated from formant structure by homomorphic filtering. A log-power periodogram is formed and then inverse-Fourier transformed to give a cepstrum containing formant data in its lower coefficients, with pitch being apparent at the higher end;

$$\text{Cepstrum} = \text{DFT}^{-1}\left\{\ln\left|\text{DFT}(\text{x}(\text{n})h(n))\right|^2\right\} \tag{34}$$

In practice, it is more computationally efficient to use an inverse discrete cosine transform (DCT) than an inverse DFT, since the periodogram is, by definition, real and symmetric about zero frequency.

Figure 14 shows the lower 40 coefficients of conventional and auditory cepstra of the example sentence. Although most of the coefficients appear rather feint (due to the first coefficient, equivalent to the log signal power, having a wider range than the rest), it can be seen that these change smoothly and with consistent "flow". The higher coefficients (which characterise pitch) also tend to evolve smoothly in the conventional cepstrum, but are almost non-existent in the auditory version. This is because the pitch harmonics are rendered aperiodic by the non-linear frequency warping. These are not shown here, however, since none of the other representations characterise pitch, and, therefore, there is nothing to compare them with.

5.6 Linear prediction coefficients

Autoregressive modelling of speech signals can give a very concise description of the vocal tract transfer function. The results of this analysis are often presented as the coefficients of a ladder filter which can be used to predict one step ahead of the speech waveform. However, these can exhibit abrupt changes even when the speech sound is changing smoothly.

Figure 15 shows the linear prediction (LP) coefficients for the example sentence. It is difficult to identify meaningful structure in these, which only appear smooth during unvoiced sounds. There is little evidence of "migration" of features between one coefficient and another.

5.7 Reflection coefficients

Burg's method for calculating LP coefficients is based on the calculation of reflection coefficients, which can be viewed as the parameters of an acoustic-pipe model of speech production (Rabiner and Schafer, 1978). These always have values between -1 and 1 and, thus, have somewhat different numerical properties from those of standard LP (ladder) coefficients, although the patterns as shown in Figure 16 seem somewhat similar. These also tend to evolve slightly more smoothly than ladder coefficients, and do exhibit slight migration of features.

5.8 Vocal tract area functions

The shape of the acoustic pipe implied by a set of reflection coefficients (see Figure 17) can be calculated by adding successive log area ratios or, alternatively, by multiplying the area ratios and then taking the log;

$$\begin{aligned}&\text{Area of section } i = A_i = A_{i-1}\frac{1-k_i}{1+k_i} \quad ; \quad 1 \le i \le p,\ A_0 = 1\\&\text{Log area of section } i = \ln\left(\prod_{m=1}^{i}\frac{1-k_m}{1+k_m}\right)\end{aligned} \tag{35}$$

This gives a set of parameters which are loosely related to the cross-sectional area of the vocal tract and, therefore, obey rules of motion similar to those of the real vocal tract. For example, as the tongue moves a constriction forward and backward, values of the vocal tract area function will move within the data vector, while the opening and closing of the mouth will affect the magnitude of the values at the respective end of that vector.

The representation in Figure 18 is, again, rather feint, because of the range of values encountered between voiced and unvoiced segments. However, this is much smoother than either reflection coefficients or LP coefficients, and does exhibit noticeable migration effects.

6 Results

The pre-processors discussed in section 5 were implemented and used to process the sentences, as described above. Mean error scores were computed for the male and the female versions of the sentence. The results are shown in a graphical form in Figure 19.

6.1 Comparison of representations

There is relatively little difference between the results for male and female speech. However, there are differences between the different representations. In particular, LP coefficients, reflection coefficients and log vocal tract area functions are all clearly difficult to predict, with or without the new flow-based techniques. This is because of the relative unimportance of "feature migration" in the overall scheme of their variability.

As expected, periodograms are also difficult to predict, presumably because of the independent changes in formants and in pitch. Lower resolution PSD estimates, however, are significantly easier to predict (including the auditory periodogram, as well as Blackman-Tukey, maximum entropy, maximum likelihood and their respective auditory versions).

The cepstral representations are by far the easiest to predict, which suggests the reason for their wide acceptance as the "best" pre-processor for ASR may lie in the appropriateness of delta and delta-delta coefficients to the characterisation of speech dynamics in the cepstral domain.

6.2 Comparison of predictors

When it comes to comparing the different prediction/interpolation algorithms, it is clear that, despite better performance when predicting steadily evolving voiced sounds (Baghai-Ravary et. al, 1994), the average performance of FBP is no better than that of the zero-order, delta coefficient, predictor. This can be attributed to the lack of robustness in the predictor, which can produce very large errors during unvoiced sounds and during abrupt transitions. This also explains why the first-order, delta-delta coefficient, predictor seems to behave worse than the zero-order version. The FBP has previously been shown to give lower errors than conventional methods during periods of smooth evolution (Baghai-Ravary et. al, 1994).

By comparison, the AFI approach gives consistently better results than the other techniques. On PSD representations, AFI is typically 25-30% more accurate. AFI is also noticeably superior in the cepstral domain, although its performance on the auditory cepstrum is somewhat less impressive. The only cases where AFI is definitely worse than the other methods, are for linear prediction and reflection coefficients. As mentioned before, these show little evidence of migratory features, so this is to be expected. However, they are also relatively unpredictable with the other techniques, so it is obviously more difficult to capture the dynamic structure of these representations.

7 Conclusion

Two new flow-based methods of modelling the speech, namely the FBP and AFI, have been presented and discussed in comparison to several other methods. Although evidenced that the FBP gives lower errors than the conventional methods during periods of smooth evolution, when used for a complete sentence, however, its lack of robustness makes its average performance no better than a zero-order (delta coefficient) predictor. AFI, however, has been shown to be an accurate predictor of smoothly-changing segments of speech as well as being robust in the presence of rapid, or unpredictable, changes. It consistently out-performs all the other methods considered. The advantage it gives is most noticeable on those PSD estimates which yield data containing a small number of relatively discrete features, such as the MEM and MLM PSD estimates which provide a combination of good frequency resolution and complete suppression of pitch information.

8 References

BAGHAI-RAVARY, L., BEET, S. W. and TOKHI, M. O. (1994). "Removing redundancy from some common representations of speech", Proceedings of the Institute of Acoustics, **16**. (to appear).

BANKS, S. P. (1990). "Signal processing, image processing and pattern recognition", Prentice Hall, Hertfordshire.

DELLER, J. R. Jr., PROAKIS, J. G. and HANSEN, J. H. L. (1993). "Discrete-Time Processing of Speech Signals", Macmillan, London.

ELLIOTT, D. F. (Ed). (1987). "Handbook of Digital Signal Processing: Engineering applications", Academic Press, London.

FLETCHER, H. (1940). "Auditory patterns", Review of Modern Physics, **12**, pp. 47-65.

KENNY, P., LENNIG, M. and MERMELSTEIN, P. (1990). "A linear predictive HMM for vector-valued observations with applications to speech recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, **38**, (2), pp. 220-225.

MAKHOUL, J. and COSELL, L. (1976). "LPCW: an LPC vocoder with linear predictive spectral warping", Proceedings of. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 466-469.

MOORE, B. C. J. and GLASBERG, B. R. (1987). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns", Hearing Research, **28**, pp. 209-225.

MUSICUS, B. R. (1985). "Fast MLM power spectrum estimation from uniformly spaced correlations", IEEE Transactions on Acoustics, Speech and Signal Processing, **33**, (4), pp. 1333-1335.

PALIWAL, K. K. (1984). "Effect of preemphasis on vowel recognition performance", Speech Communication, **3**, pp. 101-106.

PARSONS, T. W. (1987). "Voice and speech processing", McGraw-Hill, London.

PROAKIS, J. G. and MANOLAKIS, D. G. (1992). "Digital Signal Processing: Principles, algorithms and applications", 2nd Edition, Macmillan, London.

RABINER, L. R. and SCHAFER, R. W. (1978). "Digital processing of speech signals", Prentice Hall, Hertfordshire.

SHIRAI, K. and MANO, K. (1986). "A Clustering Experiment of the Spectra and Spectral Changes of Speech to Extract Phonemic Features", Signal Processing, **10**, pp. 279-290.

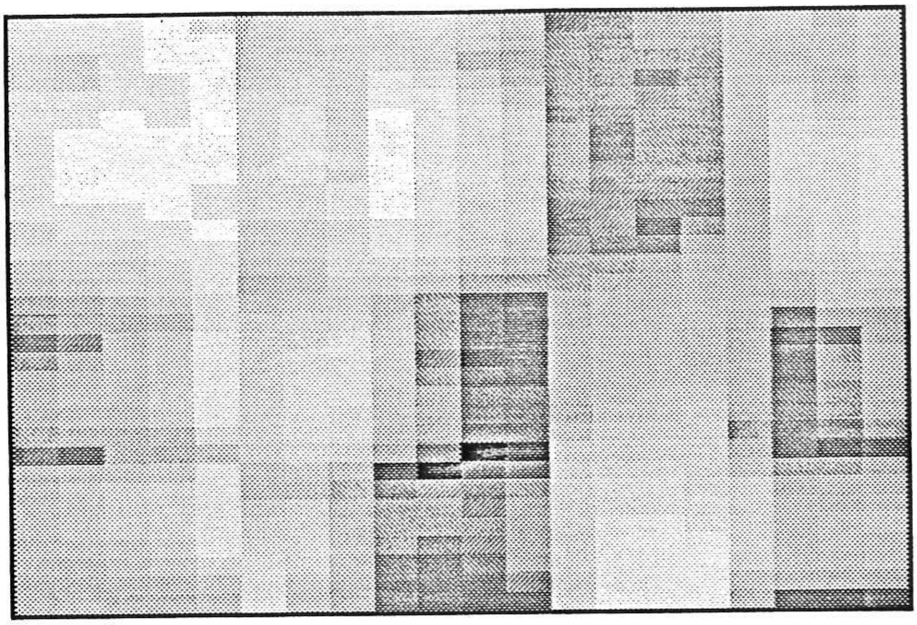


Figure 1: Spectrogram of a short phrase, "...in greasy...".

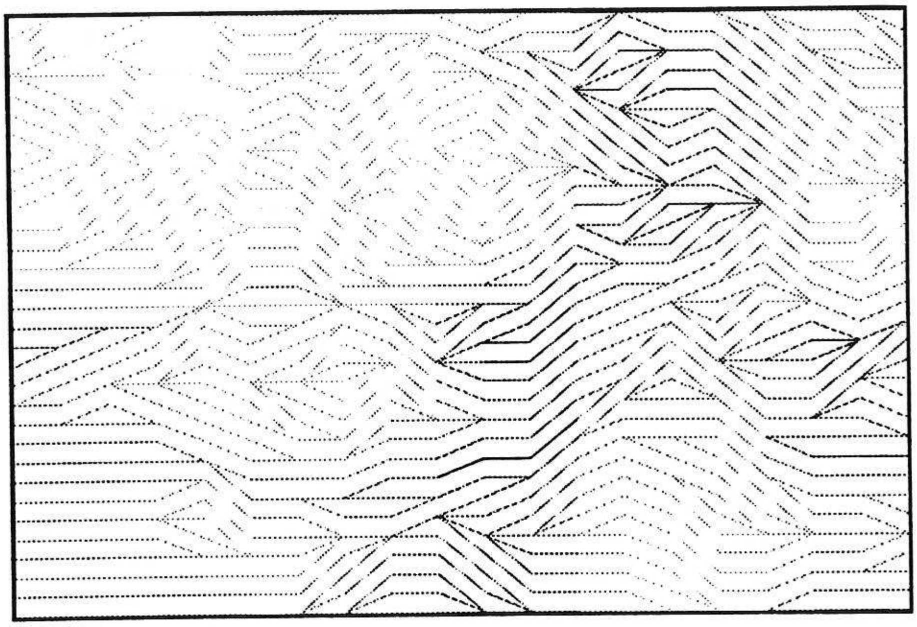


Figure 2: Spectrographic acoustic flow of a short phrase, "...in greasy...".

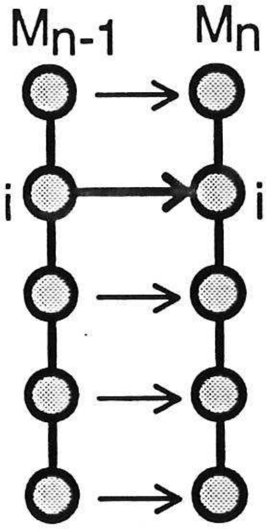


Figure 3: Zero-order linear prediction.

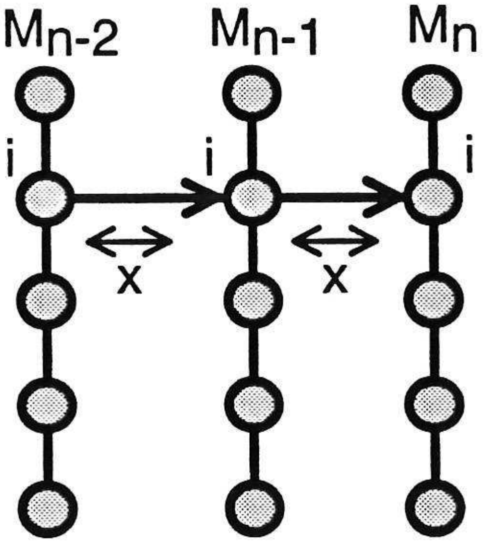


Figure 4: First-order linear prediction.

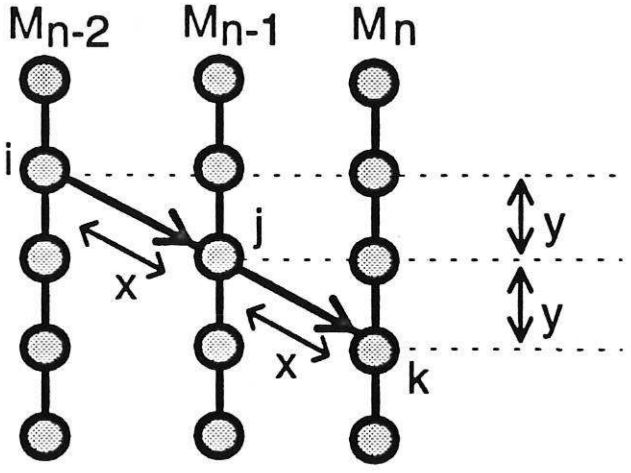


Figure 5: Flow based prediction.

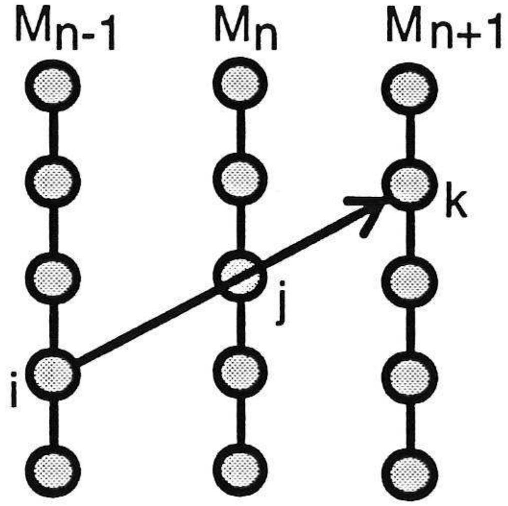


Figure 6: Acoustic flow interpolation.

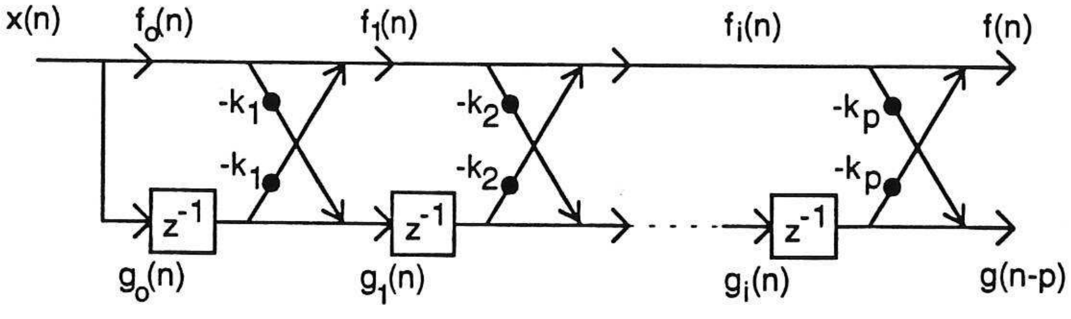


Figure 7: Lattice filter.

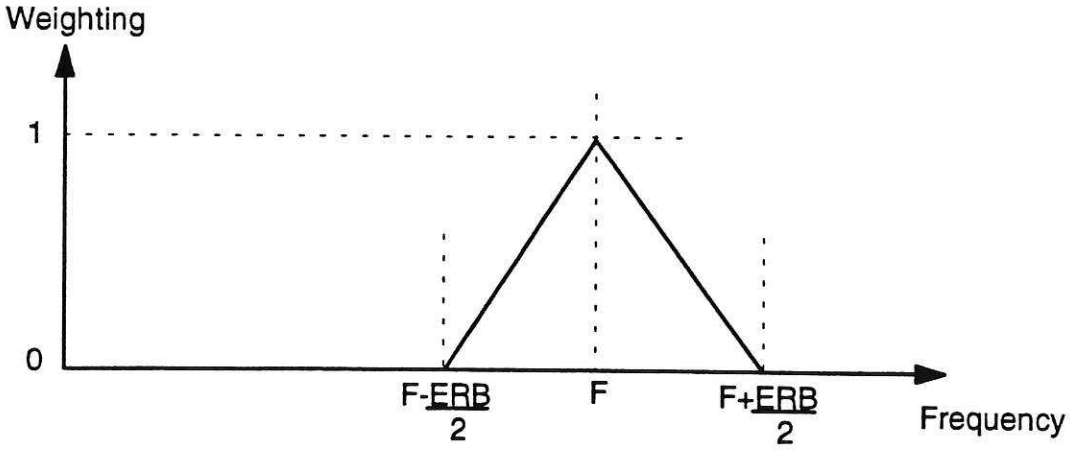
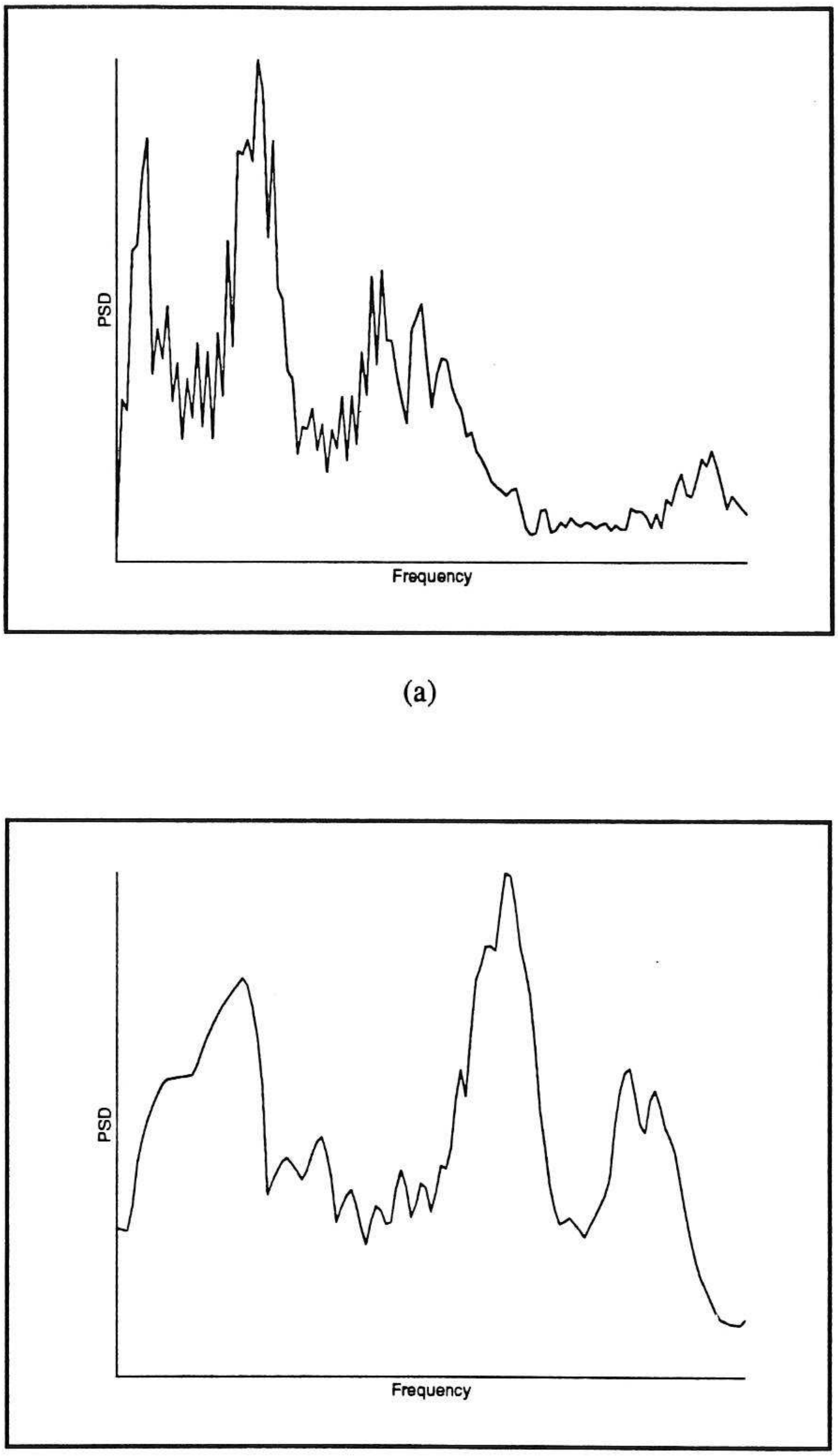


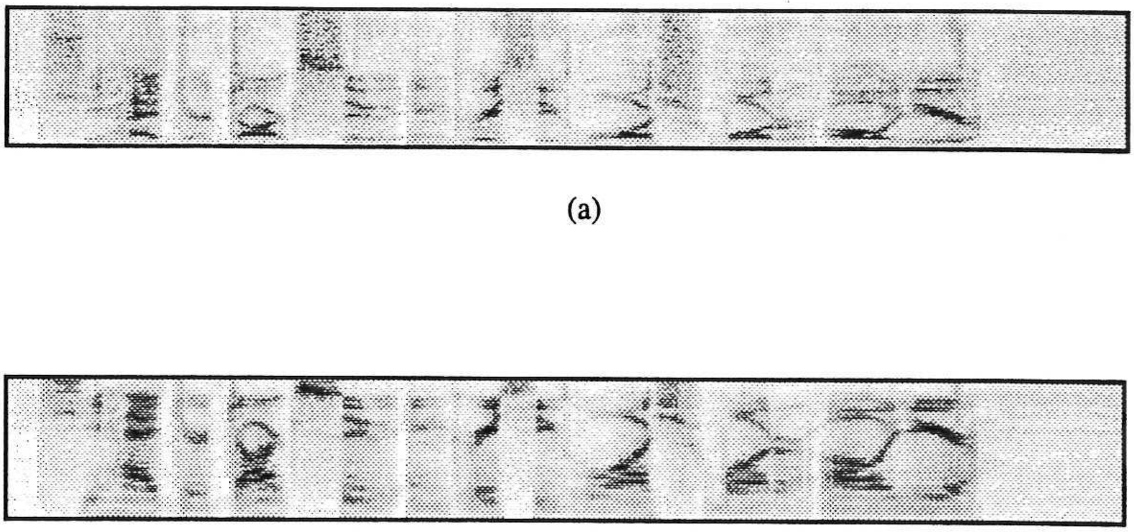
Figure 8: Triangular weighting function.



(a)

(b)

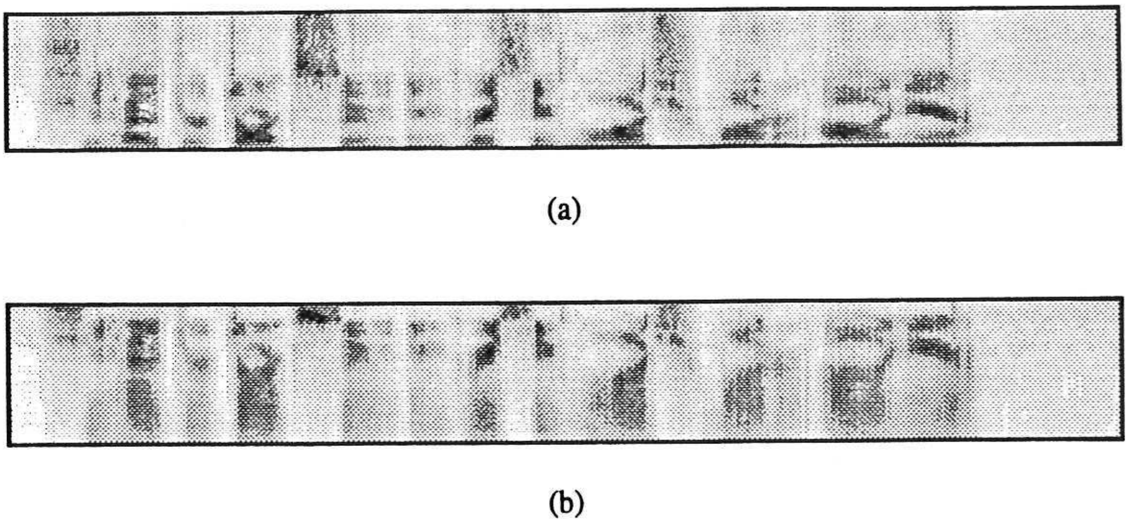
Figure 9: A voiced sound PSD value using
(a) Periodogram.
(b) Auditory periodogram.



(a)

(b)

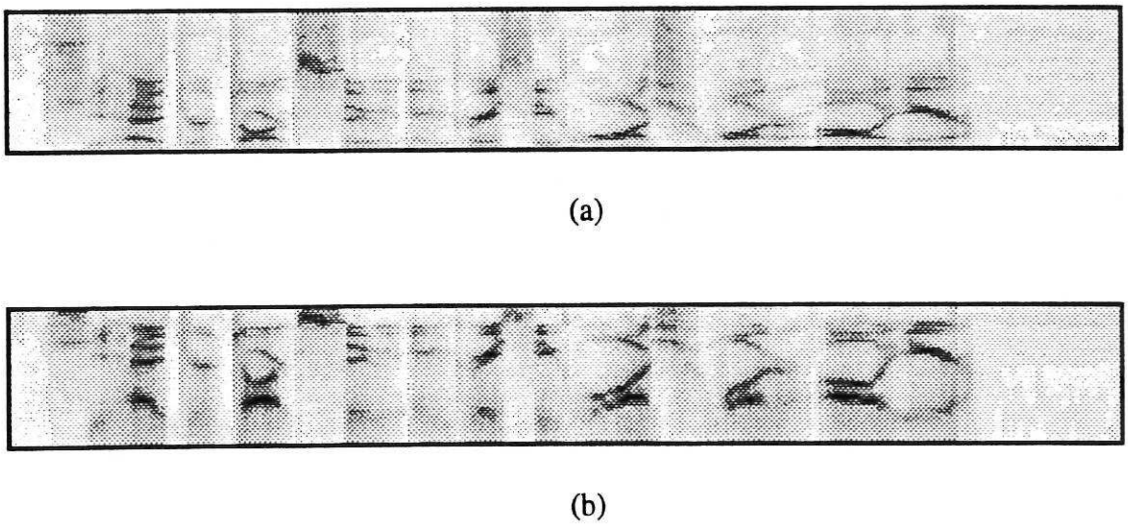
Figure 10: A sentence spoken by a male speaker, processed using
(a) Periodogram.
(b) Auditory periodogram.



(a)

(b)

Figure 11: A sentence spoken by a male speaker, processed using
(a) Blackman-Tukey PSD.
(b) Auditory Blackman-Tukey PSD.



(a)

(b)

Figure 12: A sentence spoken by a male speaker, processed using
(a) Maximum entropy PSD.
(b) Auditory maximum entropy PSD.

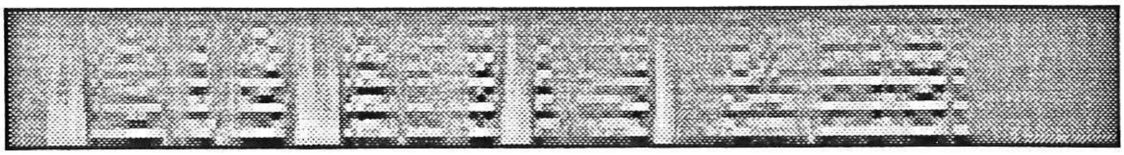


Figure 15: Linear prediction coefficients.

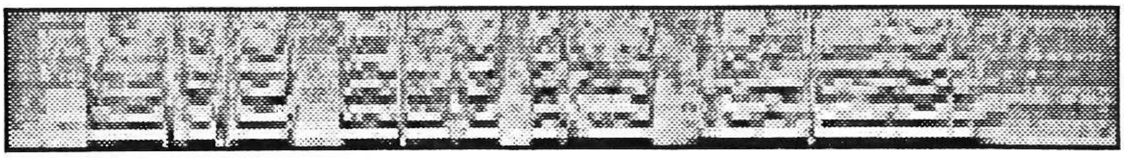


Figure 16: Reflection coefficients.

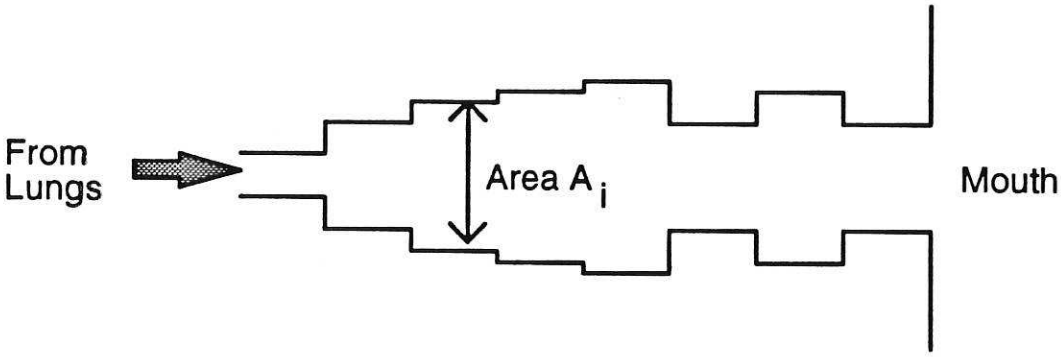


Figure 17: Acoustic pipe model of the vocal tract.

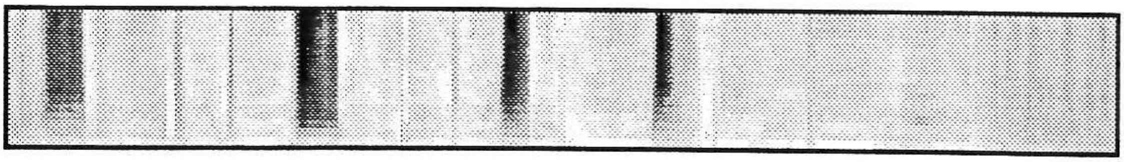
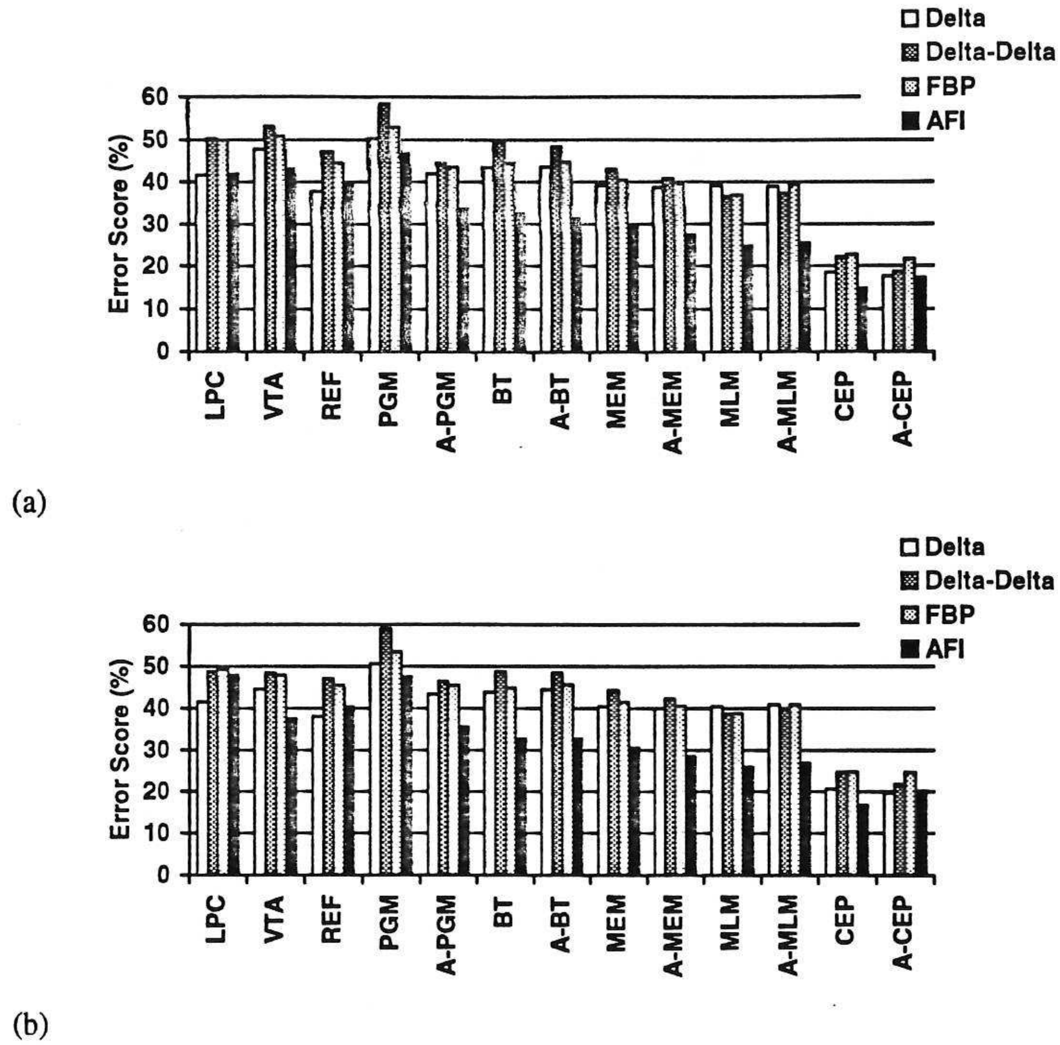


Figure 18: Vocal tract area functions.



LPC	Linear prediction (AR model) coefficients.
VTA	Log vocal tract area functions.
REF	Reflection coefficients.
PGM	Periodogram.
A-PGM	Auditory-scale periodogram.
BT	Blackman-Tukey PSD estimate.
A-BT	Auditory-scale Blackman-Tukey PSD estimate.
MEM	Maximum entropy PSD estimate.
A-MEM	Auditory-scale maximum entropy PSD estimate.
MLM	Maximum likelihood PSD estimate.
A-MLM	Auditory-scale maximum likelihood PSD estimate.
CEP	Cepstrum.
A-CEP	Auditory-scale cepstrum.

(c)

Figure 19: Mean error scores;
(a) Male version of sentence.
(b) Female version of sentence.
(c) Abbreviation of pre-processor names.